COSE474-2021F: Final Project Final Report
"Lip Reading from Video Frames: A Text Prediction Approach"

2018140088 Dosung Lee

1. Introduction

Motivation

AI services leveraging Automatic Speech Recognition (ASR) have become ubiquitous in our daily lives, seamlessly integrating into various applications and enhancing our interactive experiences. Despite the widespread integration of ASR models in our daily lives, challenges persist in accurately interpreting customers' needs, especially in the face of diverse accents, homophones, and noisy real-world environments. This project proposes a solution by incorporating lip-reading technology to augment ASR models, enhancing their precision in understanding and interpreting oral instructions.

Problem definition

While Automatic Speech Recognition (ASR) efficiently converts oral language into text, lip-reading methods take visual video frames as input and predict corresponding text for each frame. Developing a lip-reading framework with the ability to predict instructions could serve as a valuable augmentation, enabling the AI model to generate alternative responses in situations where the ASR model encounters difficulties in understanding oral instructions.

Concise description of contriubtion

This project introduces a streamlined model designed with minimal parameters to predict text from video frames, specifically addressing resource constraints such as limited GPU availability, particularly for mobile platforms. While cloud services do provide GPU resources for mobile applications, it's important to acknowledge the potential limitations such as network latency, connectivity issues, and the associated costs. Our streamlined model, optimized for minimal parameters, offers an on-device solution that not only mitigates the reliance on external servers but also ensures efficient text prediction, even in scenarios where a continuous and stable internet connection may not be guaranteed.

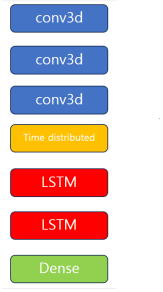

Figure 1. Illustration of the Lip Reading Framework: The model architecture incorporates three Conv3D layers with ReLU activation and max-pooling, processing the video input through 128, 256, and 75 filters, respectively. The sequence of 75 frames is maintained using a TimeDistributed layer. Subsequently, a Bidirectional LSTM, initialized with orthogonally, is applied with a dropout rate of 0.5 during training. The final Dense layer adapts the output to align with the vocabulary in the dictionary.

2. Methods

Significance novelty (main challenges and how you address them)

Our lip-reading framework draws inspiration from two state-of-the-art (SOTA) articles, seamlessly combining the strengths of each approach. The first component incorporates a structure based on Long Short-Term Memory (LSTM) and Connectionist Temporal Classification (CTC) loss, harnessing the power of sequential memory for effective speech understanding. Complementing this, we integrate a second structure derived from a Conv3D architecture, chosen specifically for its renowned capabilities in video processing. The attention layer employed in the SOTA article proved to be computationally heavy, posing challenges for our GPU resources. In response to this limitation, our framework strategically opts for an alternative solution: a dense layer with He initializer.

Reproducibility formulation

Let F represent the set of frames in the video se-

quence v, where v is the video and a is the alignment alphabet.
Each individual frame $v(i)$ corresponds to a single frame in the video sequence, where i ranges from 0 to 74.
The video frames undergo a preprocessing function $f()$ that involves cropping the frames to the region of interest: $v(i) = Crop(v(i))$ with the specific crop defined as $v(i)[190:236, 80:220, :]$.
Normalization of the frames is performed as follows:

$$mean = \frac{1}{75}\sum_{k=0}^{74} v(k)$$

$$std = \sqrt{\frac{1}{75}\sum_{k=0}^{74}(v(k) - mean)^2}$$

$$normalizedframe(i) = \frac{(v(i) - mean)}{std}$$

Here: - $v(m)$ represents the mean of all frames in the video sequence. - std is the standard deviation calculated across all frames in the video sequence. - $v(i)$ represents an individual frame in the sequence, and the normalized frame $v(i)$ is obtained by subtracting the mean and dividing by the standard deviation.

3. Experiments

Dataset

The Grid Corpus is a comprehensive dataset tailored for computational-behavioral studies in speech perception. It comprises top-quality audio and facial video recordings featuring 34 talkers (18 male, 16 female), each presenting 1000 sentences, summing up to a total of 34,000 sentences. The sentences adhere to the format 'put red at G9 now.' For the project, data from the s1 male talker was chosen. This dataset includes videos processed into 75 frames, accompanied by alignment labels assigning an alphabet to each frame. The pre-processed s1 dataset of 1000 underwent a random shuffle selection of 500, and a split resulting in 450 samples for training and 50 for testing. Detailed feature extraction procedures are outlined in the experiment settings.

Computing resources

CPU (11th Gen Intel® Core™ i5-1135G7)
16GB RAM
GPU (Colab Pro T4 GPU)
OS (Windows)
TensorFlow 2.10.1.

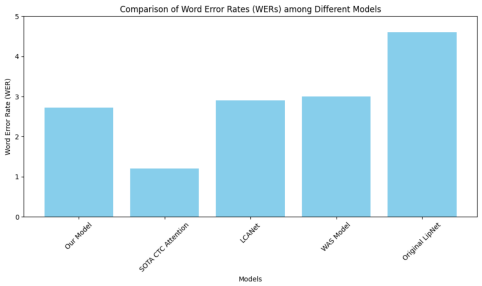

Figure 2. Quantitative results comparison with baselines, SOTA. When comparing quantitative results with baselines and the current state-of-the-art (SOTA), the CTC attention model demonstrates superior performance over all other models. Our model consistently outperforms alternative approaches, except for the CTC attention model, even when employing the simplest structure.

Experimental design/setup

To maximize computing efficiency, we conducted feature extraction on lip frames. The video processing algorithm, adapted from the original article and implemented using OpenCV, was tailored for this project. Notably, our method for mouth extraction involves a streamlined approach—simply cropping the lower portion of the video frame—departing from the baseline paper that relied on object detection. In our implementation, the RGB image underwent conversion to a Grayscale image and normalization to ensure compatibility with the input size. Regarding the learning rate scheduling, both cosine decay and exponential methods were tested, with the simpler exponential scheduler ultimately chosen due to marginal differences observed. For our model, we employed the CTC loss available in Keras due to the utilization of LSTM layers. The Keras CTC decoder was then utilized to interpret the predicted output. The Adam optimizer was chosen with an initial learning rate set at 0.0001. The Word Error Rate (WER) is calculated using the formula:

$$WER = \frac{S + I + D}{N}$$

Where: S : Number of substitutions
I : Number of insertions
D : Number of deletions
N : Total number of words in the reference text Lower WER values indicate better performance, as it means that the system's output is closer to the reference text.

4. Results

Quantitative results comparison with baselines, SOTA

The results are compared it figure 2. Our model's evaluation yielded a Word Error Rate (WER) of 2.72. Comparatively,

the SOTA CTC attention model scored a WER of 1.200, while the LCANet achieved 2.900. Additionally, the WAS model performed at a WER of 3.000, and the original LipNet demonstrated a WER of 4.600.

5. Limitations

Comparing our model to the state-of-the-art CTC attention model qualitatively was hindered by resource limitations on GPU processing for extensive datasets. Additionally, our model exhibited limited generalization capabilities when handling diverse videos, preventing a comprehensive assessment. Improving the mouth extraction method within our model is imperative to enhance its adaptability across various video types for meaningful comparisons.

6. Discussion

Our proposed CNN-based method demonstrated accurate text prediction. However, our model employed a simple technique by cropping stable video frames of the mouth, whereas other models detected and extracted features from the mouth region. This distinction might account for the observed performance of our model. While our model performed well with consistent video formats, it struggled when speakers moved significantly, resulting in inaccuracies due to challenges in accurately extracting the mouth region. This limitation contributed to higher error rates under conditions of substantial speaker movement.

In future endeavors, enhancing the precision of mouth detection and cropping will be pivotal for handling real-world scenarios more accurately. Moreover, situations may arise where the mouth isn't visible or video resolution lacks stability in uncontrolled environments. Addressing these challenges could involve a holistic approach, such as simultaneously processing audio and visual features. This joint processing strategy might serve as a key solution, allowing for more robust handling of unpredictable scenarios.

References

Assael, Y. M., Shillingford, B., Whiteson, S., & de Freitas, N. (2016). LipNet: End-to-End Sentence-level Lipreading. *arXiv [Cs.LG]*. Retrieved from http://arxiv.org/abs/1611.01599

Chung, J. S., & Zisserman, A. (2016). Lip Reading in the Wild. *Asian Conference on Computer Vision*.

Cooke, M., Barker, J., Cunningham, S., & Shao, X. (2020, January 24). *The grid audio-visual speech corpus*. Zenodo. https://zenodo.org/doi/10.5281/zenodo.3625686

He, K., Zhang, X., Ren, S., amp; Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on ImageNet Classification. 2015 IEEE International Conference on Computer Vision (ICCV).https://doi.org/10.1109/iccv.2015.123

Ma, P., Stavros, P., amp; Pantic, M. (2022). Visual Speech Recognition for Multiple Languages in the Wild. https://doi.org/10.21203/rs.3.rs-1386805/v1

Raghavendra, M., Omprakash, P., & B R, M. (2021). AuthNet: A Deep Learning Based Authentication Mechanism Using Temporal Facial Feature Movements (Student Abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(18), 15873–15874. Retrieved from https://ojs.aaai.org/index.php/AAAI/article/view/17933

Xu, K., Li, D., Cassimatis, N., amp; Wang, X. (2018). LCANet: End-to-end lipreading with cascaded attention-CTC. 2018 13th IEEE International Conference on Automatic Face amp;amp; Gesture Recognition (FG 2018). https://doi.org/10.1109/fg.2018.00088